SU UN TEOREMA CONCERNENTE LE COSIDDETTE STATISTICHE SUFFICIENTI

J. NEYMAN.

SUNTO. — L'A., riferendosi al concetto delle cosiddette «statistiche» o «valutazioni di un parametro» introdotte da R. A. Fisher, dà le condizioni necessarie e sufficienti perchè tali «valutazioni» siano «sufficienti» in un senso che è stato indicato dal Fisher stesso. Queste condizioni sono tali che è facile accertare se esse sono, oppur no, soddisfatte nei casi particolari. Pertanto esse possono essere usate come un criterio per riconoscere se una statistica sia sufficiente. La teoria è illustrata con un esempio.

1. Alcuni anni fa R. A. Fisher espresse il desiderio che si collaborasse per la dimostrazione della proposizione seguente [1]: *Se esistono delle «statistiche sufficienti» esse possono essere determinate col metodo della «massima verosimiglianza»* (maximum Likelihood).

Ci proponiamo di far vedere, in questa Nota, che la dimostrazione di questa proposizione può essere facilmente raggiunta quando le condizioni del teorema siano state chiaramente specificate.

In relazione al quesito proposto dimostreremo precisamente che, se esiste una valutazione sufficiente di un parametro, il valore che si ottiene per esso mediante il cosiddetto metodo della «massima verosimiglianza» è funzione soltanto di quella valutazione.

Cominceremo con lo spiegare la notazione da usarsi e col richiamare talune definizioni e proprietà sulle variabili casuali.

2. Indicheremo con $P\{E\}$ la probabilità che si verifichi un evento E e con $P\{E/E_1\}$ la probabilità che si verifichi lo stesso evento nell'ipotesi che un altro evento E_1 si sia già verificato. Ciò esprimeremo in seguito dicendo che $P\{E/E_1\}$ è la probabilità di E dato E_1.

---

[1] R. A. FISHER, *On the Mathematical Foundations of Theoretical Statistics.* «Philosophical Transaction of the Royal Society of London», Series A, vol. 222, pag. 323, London, 1922.

Indicando con X una variabile casuale, diremo che essa è continua se, preso un numero qualunque α, si abbia $P\{X = \alpha\} = 0$ mentre per almeno una coppia di numeri a e b con $a < b$, si abbia $P\{a < X < b\} > 0$. Diremo invece che la variabile casuale X è discontinua, se esiste un certo insieme discreto I di numeri tale che sia $P\{X = \beta\} > 0$, se β è un numero appartenente ad I e $P\{c < X < d\} = 0$ se c e d sono due numeri tali che fra essi non vi siano compresi numeri appartenenti ad I.

In ciò che segue considereremo soltanto variabili casuali delle due specie ora definite.

Consideriamo ora n variabili casuali $X_1, X_2, \cdots, X_n$ e indichiamo con H il punto dello spazio S ad n dimensioni avente per coordinate i valori $x_1, x_2, \cdots, x_n$ che le variabili indicate possono assumere in una determinata prova od esperienza.

Supponendo che le variabili casuali $X_1, X_2, \cdots, X_n$ siano continue, indichiamo con $p(x_1, x_2, \cdots, x_n)$ la densità della probabilità che dette variabili assumano valori compresi tra x_1 e $x_1 + dx_1$, x_2 e $x_2 + dx_2$, $\cdots$, x_n e $x_n + dx_n$ rispettivamente. Considerata una regione qualunque ω dello spazio S la probabilità che il punto H appartenga ad ω è data allora da

$$[1] \qquad P\{H \in \omega\} = \int \cdots \int_{\omega} p(x_1, x_2, \cdots, x_n)\, dx_1\, dx_2 \cdots dx_n ,$$

dove $H \in \omega$ significa che il punto H appartiene ad ω.

La funzione $p(x_1, x_2, \cdots, x_n)$ è definita non negativa e continua quasi dappertutto in S (eccettuato cioè al più in un insieme di punti di misura nulla). Evidentemente si ha

$$[2] \qquad \int \cdots \int_{S} p(x_1, x_2, \cdots, x_n)\, dx_1\, dx_2 \cdots dx_n = 1 ,$$

$$[3] \qquad p(x_1, \cdots, x_{n-1}) = \int_{-\infty}^{\infty} p(x_1, \cdots, x_n)\, dx_n .$$

Consideriamo ora n nuove variabili casuali continue $Y_1, Y_2, \cdots, Y_n$ e indichiamo con $p(y_1, y_2, \cdots, y_n)$ la densità della probabilità che dette variabili assumano valori rispettivamente compresi tra y_1 e $y_1 + dy_1$, y_2 e $y_2 + dy_2$, $\cdots$, y_n e $y_n + dy_n$.

Allora, se

[4] $$x_i = x_i(y_1, y_2, \cdots, y_n), \qquad (i = 1, 2, \cdots, n)$$

rappresenta una trasformazione biunivoca dello spazio S' delle y nello spazio S delle x, eccettuato al più per un insieme di punti di misura nulla; se inoltre esistono e sono continue le derivate parziali delle [4] e se il determinante funzionale

[5] $$\Delta = \frac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)}$$

non cambia segno ed è diverso da zero in quasi tutto lo spazio S', si ha

[6] $$p(y_1, \cdots, y_n) = p(x_1, \cdots, x_n)\,|\Delta|.$$

Nel seguito dovremo considerare delle funzioni $T_1, T_2, \cdots$ dei valori $x_1, x_2, \cdots, x_n$ che possono assumere le variabili casuali X_1, $X_2, \cdots, X_n$. Se queste variabili sono discontinue basterà supporre che le funzioni T siano univoche. Se invece le variabili sono continue, sarà richiesta inoltre l'esistenza e la continuità delle derivate parziali $\frac{\partial T}{\partial x_i}$ $(i = 1, 2, \cdots, n)$ rispetto a ciascun valore di x_i. D'altra parte dev'esser soddisfatta la condizione che almeno una di queste derivate parziali relative a ciascuna funzione non si annulli, salvo al più in un insieme di punti di misura nulla. Poichè l'ordine nel quale le variabili vengono considerate è arbitrario, supporremo $\frac{\partial T}{\partial x_1} \neq 0$.

Se si considerano due funzioni indipendenti, T_1, T_2, si supporrà similmente che $\frac{\partial(T_1, T_2)}{\partial(x_1, x_2)}$ non si annulli salvo al più in un insieme di punti di misura nulla.

Consideriamo di nuovo le variabili casuali $X_1, X_2, \cdots, X_n$ e indichiamo con $p(x_{k+1}, \cdots, x_n / x_1^0, \cdots, x_k^0)$ la densità della probabilità che le variabili $X_{k+1}, \cdots, X_n$ assumano rispettivamente valori compresi tra x_{k+1} e $x_{k+1} + dx_{k+1}, \cdots, x_n$ e $x_n + dx_n$, nell'ipotesi che le rimanenti variabili abbiano assunto valori fissi $x_1^0, \cdots, x_k^0$.

Allora, qualunque sia la regione ω dello spazio ad $n - k$ dimensioni e qualunque siano i valori fissi $x_1^0, \cdots, x_k^0$ avremo, indicando con H un punto di questo spazio di coordinate $x_{k+1}, \cdots, x_n$

$$[7] \qquad P\{(H \in \omega)/(X_1 = x_1^\circ), \cdots, (X_k = x_k^\circ)\} =$$

$$= \int \cdots \int_\omega p(x_{k+1}, \cdots, x_n / x_1^\circ, \cdots, x_k^\circ)\, dx_{k+1} \cdots dx_n .$$

Si può inoltre dimostrare che

$$[8] \qquad p(x_1, \cdots, x_k, x_{k+1}, \cdots, x_n) =$$

$$= p(x_1, \cdots, x_k) \cdot p(x_{k+1}, x_{k+2}, \cdots, x_n / x_1, x_2, \cdots, x_k).$$

Osserviamo infine che quando due densità di probabilità differiscono soltanto in un insieme di punti di misura nulla, esse dovranno essere considerate come identiche.

Faremo uso di questa osservazione più avanti.

Nel caso in cui le variabili casuali $X_1, \cdots, X_n$ siano discontinue, indicheremo con $p(x_1, \cdots, x_n)$ la probabilità che esse assumano rispettivamente i valori $x_1, \cdots, x_n$.

Si ha cioè

$$[9] \qquad p(x_1, \cdots, x_n) = P\{(X_1 = x_1), \cdots, (X_n = x_n)\} .$$

Tutte le formule precedenti riguardanti variabili continue si possono estendere in modo evidente al caso di variabili discontinue. Naturalmente dove avevamo degli integrali si avranno delle somme di valori di $p(x_1, \cdots, x_n)$ estese a determinati possibili valori delle x.

3. Supponiamo che la probabilità che le variabili $X_1, \cdots, X_n$ assumano rispettivamente valori compresi tra x_1 e $x_1 + dx_1, \cdots, x_n$ e $x_n + dx_n$, oppure assumano i valori $x_1, \cdots, x_n$, secondochè le variabili stesse siano continue o discontinue, dipenda da un parametro a il cui valore sia incognito.

Indichiamo con T delle funzioni delle x non contenenti a. Alcune di queste funzioni T sono tali che, sostituendo in esse alle x dei particolari valori osservati x°, i valori che tali funzioni vengono ad assumere possono considerarsi quali valori approssimati del parametro a.

Le funzioni utili per tali scopi sono chiamate « valutazioni di a » o brevemente « statistiche ».

Per esempio, se la densità della probabilità relativa ad n variabili casuali continue $X_1, X_2, \cdots, X_n$ è data da

$$[10] \qquad p(x_1, x_2, \cdots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^2}},$$

dove a e σ sono due parametri incogniti, il parametro a può essere valutato mediante le seguenti funzioni

$$[11] \qquad T_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$[12] \qquad T_2 = \frac{1}{2}(x_1 + x_n),$$

dove x_1 e x_n indicano il più piccolo e il più grande valore delle x.

Evidentemente possono suggerirsi anche altre valutazioni di a.

Il parametro σ può essere valutato per mezzo della funzione

$$[13] \qquad T_3 = \frac{c}{n}\sum_{i=1}^{n} |x_i - \bar{x}|,$$

dove c è una costante, oppure per mezzo della funzione

$$[14] \qquad T_4 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Queste varie funzioni non ci daranno i valori dei parametri con la stessa approssimazione, ma finchè non si richieda una determinata approssimazione si può usare indifferentemente l'una o l'altra di esse per la valutazione dei parametri stessi.

4. Un'importante classe di « statistiche » è costituita dalle cosiddette « statistiche sufficienti » che furono considerate per la prima volta da R. A. Fisher [2].

Premettiamo che una statistica o valutazione T di un parametro a si può considerare quale valore assunto, in una prova, di una variabile casuale $\mathfrak{T}$ e possiamo indicare con $p(T)$ la probabilità che $\mathfrak{T}$

[2] Cfr. loc. cit. [1].

assuma il valore T. Ciò se la variabile $\mathfrak{T}$ è discontinua; se invece $\mathfrak{T}$ è continua, $p(T)$ indicherà la densità della probabilità che la variabile $\mathfrak{T}$ assuma un valore compreso fra T e $T + dT$.

Considerata un'altra valutazione T_1 del parametro a, e la corrispondente variabile casuale $\mathfrak{T}_1$, continua o discontinua, indicheremo con $p(T_1/T)$ la probabilità che la variabile $\mathfrak{T}_1$ assuma il valore T_1 quando la $\mathfrak{T}$ abbia assunto il valore T, se per esempio le variabili $\mathfrak{T}$ e $\mathfrak{T}_1$ sono discontinue. Nel caso che queste variabili siano continue la $p(T_1/T)$ rappresenterà la densità della corrispondente probabilità.

Ciò premesso:

Una statistica o valutazione T di un parametro a è detta sufficiente, se per ogni altra valutazione T_1 la funzione $p(T_1/T)$ è indipendente da a.

È evidente che se T è una valutazione sufficiente di a, anche qualunque funzione monotona in senso stretto di T è pure una valutazione sufficiente di a.

Più generalmente. se T è una statistica sufficiente di un parametro a e T_1 una funzione univoca di T e tale che ad un valore di T_1 corrisponda uno ed un solo valore di T, allora T_1 è pure una statistica sufficiente. D'altra parte, se T_1 è una funzione univoca di T, che assuma lo stesso valore in corrispondenza a due differenti valori di T, allora T_1 non è necessariamente una statistica sufficiente.

La ragione per cui le statistiche aventi la proprietà ora detta sono chiamate «sufficienti» apparisce da quanto segue.

Intanto è ovvio che se la legge di probabilità relativa ad una funzione delle x, ad esempio T_1, non contiene il parametro a che si tratta di stimare, non sarà possibile determinare a, in modo che, ad esempio, risulti massima la funzione $p(T_1)$, e pertanto una siffatta funzione T_1 è per noi insignificante.

Ora se T è una «statistica sufficiente» secondo la suindicata definizione, ogni altra statistica T_1 sarà insignificante poichè la legge di probabilità di T_1 nell'ipotesi che T sia conosciuta è, per definizione, indipendente da a. In altri termini se una statistica T è sufficiente per la determinazione di a, ogni altra statistica T_1 è da considerarsi inutile.

5. Il metodo che si pone a base della ricerca delle valutazioni di un parametro a, conosciuto col nome di «metodo della massima verosimiglianza», consiste nel determinare, per ogni sistema di valori

x_i $(i = 1, 2, \cdots, n)$ quel valore $\hat{a}$ del parametro a che rende massima la funzione $p(x_1, x_2, \cdots, x_n)$.

Dimostreremo un teorema che può essere enunciato come segue:

TEOREMA I. — *Se esiste una valutazione sufficiente T del parametro a, il valore $\hat{a}$ che si ottiene col metodo della massima verosimiglianza è una funzione della sola T e pertanto non dipende esplicitamente dalle x.*

La dimostrazione di questo teorema è basata sul seguente

TEOREMA II. — *Se T è una valutazione sufficiente del parametro a, allora la funzione $p(x_1, \cdots, x_n)$ può essere presentata in forma di prodotto*

$$[15] \qquad p(x_1, \cdots, x_n) \equiv p(T) \cdot \Phi(x_1, \cdots, x_n),$$

dove $p(T)$ dipende da a e $\Phi(x_1, \cdots, x_n)$ è una determinata funzione delle x non contenente a.

Inversamente, se T è una funzione delle sole x e se la [15] *è soddisfatta, allora T è una valutazione sufficiente del parametro a.*

La seconda parte del Teorema II si dimostra subito. Consideriamo per esempio il caso in cui le X_i $(i = 1, 2, \cdots, n)$ siano continue.

Supponiamo allora che la funzione T verifichi la [15] e sia T_1 una nuova funzione delle x indipendente da a, tale che sia impossibile considerarla come funzione univoca di T [3].

Poichè, per le ipotesi fatte, il determinante funzionale $\dfrac{\partial(x_1, x_2)}{\partial(T, T_1)}$ non è identicamente nullo, la densità della probabilità $p(T_1/T)$ si otterrà integrando rispetto a $x_3, x_4, \cdots, x_n$ il prodotto

$$\Phi(x_1, \cdots, x_n) \left| \frac{\partial(x_1, x_2)}{\partial(T, T_1)} \right|,$$

dove x_1 ed x_2 saranno espresse in funzione di $T, T_1, x_3, \cdots, x_n$. Essendo questo prodotto indipendente da a tale sarà pure $p(T_1/T)$ e ciò significa che T è una statistica sufficiente.

Il caso in cui le X siano discontinue si tratta in maniera perfettamente analoga.

---

3) È ovvio che non sarebbe utile considerare la funzione $T_1 = f(T)$ quale funzione univoca di T, poichè in tal caso il valore di T determina quello di T_1 ed allora $p(T_1/T)$ sarebbe sempre nulla salvo che per $T_1 = f(T)$ nel qual caso sarebbe eguale all'unità.

Prima di procedere oltre può essere utile illustrare il significato del Teorema II.

Supponiamo di nuovo che la [10] rappresenti la densità della probabilità che le variabili casuali $X_1, X_2, \cdots, X_n$ assumano valori compresi tra x_1 e $x_1 + dx_1$, x_2 e $x_2 + dx_2, \cdots, x_n$ e $x_n + dx_n$. È noto allora che la densità della probabilità che la variabile casuale

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

assuma un valore compreso fra $\bar{x}$ e $\bar{x} + d\bar{x}$ è data da

$$[16] \qquad p(\bar{x}) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{n(\bar{x}-a)^2}{2\sigma^2}}$$

Confrontando con la [10] otteniamo facilmente

$$[17] \qquad p(x_1, x_2, \cdots, x_n) \equiv p(\bar{x}) \cdot \Phi(x_1, x_2, \cdots, x_n),$$

dove

$$[18] \qquad \Phi(x_1, x_2, \cdots, x_n) = \frac{1}{\sqrt{n}} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2}}.$$

Poichè la funzione $\Phi(x_1, x_2, \cdots, x_n)$ non contiene a, segue dal Teorema II che $T_1 \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ è una valutazione sufficiente di a.

Come altro esempio è facile vedere che $S = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ non è una valutazione sufficiente di σ. Infatti la densità della probabilità che la nuova variabile casuale $\delta = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ assuma un valore compreso tra S e $S + dS$ è data per $S \geqq 0$ da

$$[19] \qquad p(S) = \frac{1}{2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^{n-1}} S^{n-2} e^{-\frac{S^2}{2\sigma^2}},$$

da cui segue

$$[20] \qquad p(x_1, x_2, \cdots, x_n) \equiv p(S) \cdot \Psi(x_1, x_2, \cdots, x_n, \sigma),$$

dove

$$[21] \qquad \Psi(x_1, x_2, \cdots, x_n, \sigma) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{2^{\frac{3}{2}} \pi^{\frac{n}{2}} \sigma} S^{-n+2} e^{-\frac{n(\bar{x}-a)^2}{2\sigma^2}} \cdot$$

Poichè la funzione $\Psi(x_1, x_2, \cdots, x_n, \sigma)$ dipende da σ ne viene che S non è una valutazione sufficiente di σ. Ne segue che non sono valutazioni sufficienti di σ nemmeno $S/\sqrt{n-1}$ ed $S/\sqrt{n}$.

6. Osserviamo ancora, prima di dimostrare la proposizione che costituisce la prima parte del Teorema II, che, una volta dimostrata tale proposizione, il Teorema I diventa un semplice corollario di essa.

Scegliendo infatti $a = \hat{a}$ in maniera da rendere massima la funzione $p(x_1, \cdots, x_n)$ per un assegnato sistema di valori delle x, sarà resa massima anche $p(T)$ essendo $\Phi(x_1, \cdots, x_n)$ indipendente da a. Allora, poichè $p(T)$ non dipende esplicitamente dalle x ma vi dipende solo attraverso la funzione T, il valore $a = \hat{a}$ che rende massima $p(T)$ dipende unicamente dal valore di T.

Abbiamo quindi soltanto da dimostrare la prima parte del Teorema II.

7. Supponiamo che la funzione $T_1(x_1, \cdots, x_n)$ sia una valutazione sufficiente di a. Consideriamo la funzione $p(T_1, y_2, \cdots, y_n)$ relativa ad un nuovo sistema di variabili casuali $\mathcal{T}_1, Y_2, \cdots, Y_n$. Le variabili casuali indicate siano legate alle X_i $(i = 1, 2, \cdots, n)$ mediante le relazioni

$$[22] \qquad \begin{aligned} \mathcal{T}_1 &= \mathcal{T}_1(X_1, \cdots, X_n), \\ Y_r &= X_r, \quad (r = 2, \cdots, n). \end{aligned}$$

Per la [6] avremo

$$[23] \qquad p(x_1, \cdots, x_n) = p(T_1, y_2, \cdots, y_n) \left| \frac{\partial T_1}{\partial x_1} \right|$$

se le variabili casuali considerate sono continue e

$$[24] \qquad p(x_1, \cdots, x_n) = p(T_1, y_2, \cdots, y_n)$$

se sono discontinue.

In entrambi i casi le uguaglianze [23] e [24] valgono sotto la condizione che la corrispondenza fra le X e le nuove variabili sia espressa mediante le [22]. Ora, siano continue o discontinue le variabili, abbiamo

$$[25] \qquad p(T_1, y_2, \cdots, y_n) = p(T_1) \cdot p(y_2, \cdots, y_n / T_1)$$

dove $p(T_1)$ dipende da a; perciò si ha

$$[26] \qquad p(x_1, \cdots, x_n) = p(T_1) \cdot p(y_2, \cdots, y_n / T_1) \left| \frac{\partial T_1}{\partial x_1} \right|,$$

nel caso in cui le variabili siano continue e

$$[27] \qquad p(x_1, \cdots, x_n) = p(T_1) \cdot p(y_2, \cdots, y_n / T_1)$$

nell'altro caso.

Poichè la funzione $T_1(x_1, \cdots, x_n)$ non dipende da a, anche la sua derivata $\frac{\partial T_1}{\partial x_1}$ non dipende da a. Perciò, in entrambi i casi, affinchè la prima parte del Teorema II resti dimostrata, avremo soltanto da provare che $p(y_2, \cdots, y_n / T_1)$ è indipendente da a.

Nel caso in cui le variabili casuali X_i $(i = 1, 2, \cdots, n)$ e di conseguenza le Y_r $(r = 2, 3, \cdots, n)$ e la $\mathfrak{T}_1$ siano discontinue, ciò può essere provato senza alcuna restrizione.

Se invece dette variabili sono continue, la indipendenza di $p(y_2, \cdots, y_n / T_1)$ dal parametro a può essere dimostrata soltanto con riferimento ai sistemi di valori $y_2, \cdots, y_n$ per i quali $p(y_2, \cdots, y_n / T_1)$ è continua.

Infatti, poichè l'insieme di punti dello spazio ad $n - 1$ dimensioni nel quale $p(y_2, \cdots, y_n / T_1)$ può essere discontinua è di misura nulla, i valori di $p(y_2, \cdots, y_n / T_1)$ in questi punti possono essere del tutto arbitrari e quindi possono dipendere da a. Pertanto, poichè le densità della probabilità che differiscono soltanto in un insieme di punti di misura nulla debbono essere considerate come identiche, avremo da esaminare i valori di $p(y_2, \cdots, y_n / T_1)$, corrispondenti a differenti valori del parametro a, soltanto in punti di continuità. Con questa restrizione possiamo dimostrare la prima parte del Teorema II.

Supponiamo che questo teorema non sia vero e indichiamo con $y_2^0, y_3^0, \cdots, y_n^0$ un sistema di valori per i quali i valori $p_1(y_2^0, \cdots, y_n^0 / T_1)$ e $p_2(y_2^0, \cdots, y_n^0 / T_1)$ assunti da $p(y_2, \cdots, y_n / T_1)$

in corrispondenza a due diversi valori di a, rispettivamente a_1 e a_2, non siano gli stessi.

Supponiamo, per esempio, che sia

$$[28] \qquad p_1(y_2^0, \cdots, y_n^0/T_1) = p_2(y_2^0, \cdots, y_n^0/T_1) + 2\alpha,$$

con $\alpha > 0$.

Come abbiamo già osservato, se le variabili $Y_r (r = 2, 3, \cdots, n)$ sono discontinue non vi saranno restrizioni per $y_2^0, \cdots, y_n^0$, mentre nel caso della continuità sarà necessario che tanto $p_1(y_2^0, \cdots, y_n^0/T_1)$ quanto $p_2(y_2^0, \cdots, y_n^0/T_1)$ siano continue nel punto $y_2^0, \cdots, y_n^0$.

Indichiamo ora con $\mathfrak{T}_2$ una nuova variabile casuale legata alle Y_r mediante la relazione

$$[29] \qquad \sum_{r=2}^{n} (Y_r - y_r^0)^2 = \mathfrak{T}_2^2.$$

Vedremo che sotto la condizione [28] la funzione $p(T_2/T_1)$ dipende da a.

Consideriamo in primo luogo il caso in cui le variabili siano continue.

Per la continuità di $p_1(y_2, \cdots, y_n/T_1)$ e $p_2(y_2, \cdots, y_n/T_1)$ nel punto di coordinate $y_2^0, \cdots, y_n^0$ possiamo determinare un valore $T_2 = T_2^0$ tale che sia

$$[30] \qquad p_1(y_2, \cdots, y_n/T_1) \geqq p_2(y_2, \cdots, y_n/T_1) + \alpha$$

entro tutta la ipersfera S_0,

$$[31] \qquad \sum_{r=2}^{n} (y_r - y_r^0)^2 = T_2^{0^2}.$$

Indichiamo con $P\{(T_2 \leqq T_2^0)/T_1, a_i\}$ la probabilità che un punto di coordinate $y_2, \cdots, y_n$ appartenga alla ipersfera S_0, quando $\mathfrak{T}_1$ assuma un valore fisso T_1 e a il valore a_i con $i = 1, 2$.

Avremo

$$[32] \quad P\{(T_2 \leqq T_2^0)/T_1, a_i\} = \int \cdots \int_{S_0} p_i(y_2, \cdots, y_n/T_1) dy_2 \cdots dy_n, \ (i = 1, 2)$$

e per la [30]

$$[33] \qquad P\{(T_2 \leqq T_2^0)/T_1, a_1\} > P\{(T_2 \leqq T_2^0/T_1, a_2\}.$$

Ma $P\{(T_2 \leq T_2^\circ)/T_1, a_i\}$ può anche essere interpretata come la probabilità, data T_1, che $\mathfrak{T}_2$ assuma un valore non superiore a T_2°. Conseguentemente

$$[34] \qquad P\{(T_2 \leq T_2^\circ)/T_1, a_i\} = \int_0^{T_2^\circ} p_i(T_2/T_1)\, dT_2, \qquad (i = 1, 2)$$

dove $p_i(T_2/T_1)$ indica la densità della probabilità di $\mathfrak{T}_2$ corrispondente al valore di $\mathfrak{T}_1$ precedentemente fissato e al valore del parametro $a = a_i$ $(i = 1, 2)$.

Allora, perchè la [33] sia valida, dovrà esistere un insieme di valori di T_2 di misura non nulla nel quale sia

$$[35] \qquad p_1(T_2/T_1) > p_2(T_2/T_1);$$

ne segue quindi che $p(T_2/T_1)$ dipende da a.

Questa conclusione contraddice all'ipotesi fatta che T_1 sia una statistica sufficiente di a e quindi tale che qualunque sia la funzione T_2 la funzione $p(T_2/T_1)$ risulti indipendente da a. Ne viene quindi che $p(y_2, \cdots, y_n/T_1)$ è indipendente da a.

Nel caso in cui le variabili siano discontinue la dimostrazione è ancora più semplice.

Infatti, se vale la [28], possiamo subito concludere che $p(T_2/T_1)$ dipende da a, poichè per $T_2 = 0$ abbiamo

$$[36] \qquad p_i(T_2/T_1) = p_i(y_2^\circ, \cdots, y_n^\circ/T_1).$$

Così il Teorema II e quindi il Teorema I sono dimostrati.

Di solito il metodo della massima verosimiglianza è applicato nel caso in cui la funzione $p(x_1, \cdots, x_n)$ sia differenziabile rispetto ad a.

Allora il valore $\hat{a}$ di a che si ottiene col metodo della massima verosimiglianza si ha risolvendo l'equazione

$$[37] \qquad \frac{\partial(\log p)}{\partial a} = 0.$$

La dimostrazione precedente, però, non dipende dall'ipotesi che la funzione di a, $p(x_1, \cdots, x_n)$, sia differenziabile rispetto ad a o anche soltanto continua.

In certi problemi di genetica si hanno da considerare dei parametri che possono assumere soltanto valori interi. Anche in questi

casi il metodo della massima verosimiglianza può essere applicato con buon esito per trovare, quando esista, una « valutazione sufficiente » di tali parametri.

8. Sarà utile illustrare con un esempio l'uso del teorema dimostrato.

Nell'esempio che daremo il parametro a non è una variabile continua e pertanto non dà luogo a differenziazione.

Supponiamo che un certo carattere ereditario $\mathfrak{C}$ di un organismo sia un carattere recessivo dipendente da un numero incognito di coppie di geni. Indicheremo questi geni con a_i (recessivi) ed A_i (dominanti) $(i = 1, 2, \cdots, n)$.

Ne segue che un individuo possederà il carattere $\mathfrak{C}$ se la sua composizione genetica è data da

$$[38] \qquad a_1, a_1; a_2, a_2; \cdots; a_n, a_n,$$

e non possederà il carattere $\mathfrak{C}$ in tutti gli altri casi.

Per valutare il numero incognito n delle coppie di geni aventi influenza su $\mathfrak{C}$ supponiamo siano stati fatti parecchi incroci fra gli ibridi, fra individui, cioè, aventi la seguente composizione genetica

$$[39] \qquad A_1, a_1; A_2, a_2; \cdots; A_n, a_n.$$

Indichiamo con k il numero degli incroci fatti, con m_i il numero dei nati dall'i^{mo} incrocio e con x_i il numero dei nati dall'i^{mo} incrocio che hanno il carattere $\mathfrak{C}$.

I valori x_i $(i = 1, 2, \cdots, k)$ possono pensarsi quali valori assunti da k variabili casuali X_i $(i = 1, 2, \cdots, k)$.

In accordo con la teoria di Mendel, la probabilità che un nato assegnato abbia il carattere $\mathfrak{C}$ è uguale a $4^{-n} = p_n$. La probabilità che fra m_i nati ve ne siano esattamente x_i aventi il carattere $\mathfrak{C}$ è pertanto

$$\frac{m_i!}{x_i!\,(m_i - x_i)!}\, p_n^{x_i}\,(1 - p_n)^{m_i - x_i}, \quad (i = 1, 2, \cdots, k).$$

Infine, la probabilità che le k variabili $X_1, X_2, \cdots, X_k$ assumano i valori $x_1, x_2, \cdots, x_k$ è data da

$$[40] \qquad p(x_1, x_2, \cdots, x_k) = \prod_{i=1}^{k} \frac{m_i!}{x_i\,(m_i - x_i)!}\, p_n^{x_i}\,(1 - p_n)^{m_i - x_i}.$$

Vediamo ora se vi è una funzione delle x che costituisca una « valutazione sufficiente » del parametro incognito n.

Applichiamo per questo il Teorema I e determiniamo una valutazione di n, col metodo della massima verosimiglianza: se esiste una valutazione sufficiente di n, il valore trovato col metodo della massima verosimiglianza sarà una funzione di quella valutazione sufficiente.

Indichiamo con $\hat{n}$ il valore di n che rende massima la [40] o il più piccolo dei valori che godono di questa proprietà, se ve ne è più d'uno. Avremo allora

$$[41]\quad p(x_1, \cdots, x_k/\hat{n}-1) < p(x_1, \cdots, x_k/\hat{n}) \geqq p(x_1, \cdots, x_k/\hat{n}+1),$$

ovvero, per la [40],

$$[42]\quad p_{\hat{n}-1}^{x}(1-p_{\hat{n}-1})^{m-x} < p_{\hat{n}}^{x}(1-p_{\hat{n}})^{m-x} \geqq p_{\hat{n}+1}^{x}(1-p_{\hat{n}+1})^{m-x},$$

dove

$$[43]\qquad x = \sum_{i=1}^{k} x_i \quad , \quad m = \sum_{i=1}^{k} m_i .$$

Si vede subito che se m ha un valore fisso allora $\hat{n}$ è funzione della sola x. Perciò, in tal caso, se esistono valutazioni sufficienti di n la x deve essere una di esse, ogni altra essendo funzione di x.

Dobbiamo ora vedere se x soddisfa alle condizioni sufficienti per essere una « valutazione sufficiente ». Queste condizioni sono date dal Teorema II.

Per applicare questo teorema scriviamo la espressione che dà la probabilità che la variabile casuale $X = \sum_{i=1}^{k} X_i$ assuma il valore x.

Essa è data da

$$[44]\qquad p(x) = \frac{m!}{x!\,(m-x)!}\, p_n^{x}(1-p_n)^{m-x}.$$

Confrontando questa eguaglianza con la [40] troviamo facilmente

$$[45]\qquad p(x_1, x_2, \cdots, x_k/n) \equiv p(x) \cdot \Phi(x_1, x_2, \cdots, x_k),$$

dove

$$[46] \qquad \Phi(x_1, x_2, \cdots, x_k) = \frac{x!(m-x)!}{m!} \prod_{i=1}^{k} \frac{m_i!}{x_i!(m_i-x_i)!}.$$

Poichè la funzione $\Phi(x_1, x_2, \cdots, x_k)$ non dipende da n, si può concludere che x è una valutazione sufficiente di n.

È ovvio tuttavia che x da sè stesso non può essere considerato come un valore approssimato di n. Un tal valore potrebbe essere fornito da una funzione di x che dipenda dal metodo di valutazione che conviene applicare.

Per esempio, potremo usare il valore di n ottenuto col metodo della massima verosimiglianza, risolvendo la [42] rispetto ad $\hat{n}$; si ha

$$\frac{\log(4^{Q+1}-1) - \log(4^Q-1)}{\log 4} - 1 \leq \hat{n} < \frac{\log(4^{Q+1}-1)\log(4^Q-1)}{\log 4},$$

dove $Q = x/(m-x)$.

D'altra parte potremmo anche applicare il principio della speranza matematica. In tal caso dovremmo considerare la speranza matematica di X, cioè $E(X) = m \cdot 4^{-n}$ e considerare il valore $n = \bar{n}$ per il quale $E(X)$ eguagli la media aritmetica dei valori osservati di X. Il parametro n sarebbe allora valutato mediante l'intero più vicino a $\bar{n}$.

Le due valutazioni non saranno necessariamente eguali.

Qualunque possa essere il metodo di stima, nel caso presente, segue però che noi possiamo fermare la nostra attenzione su x e considerare una funzione della sola x come quella che possa fornire un valore approssimato di n. Una volta nota x i singoli valori di $x_1, x_2, \ldots, x_k$ o di una loro funzione non ci daranno nessun nuovo ragguaglio su n.

I teoremi I e II sono stati dimostrati sotto le ipotesi restrittive concernenti la regolarità della statistica T nel caso che le variabili casuali $X_1, X_2, \cdots, X_n$ siano continue. I casi considerati sono i più semplici e, nello stesso tempo, i più importanti dal punto di vista delle applicazioni. Comunque non vi è dubbio che il campo di applicazione dei teoremi sia molto più vasto.